Worksheet 1 Compression and encryption Answers

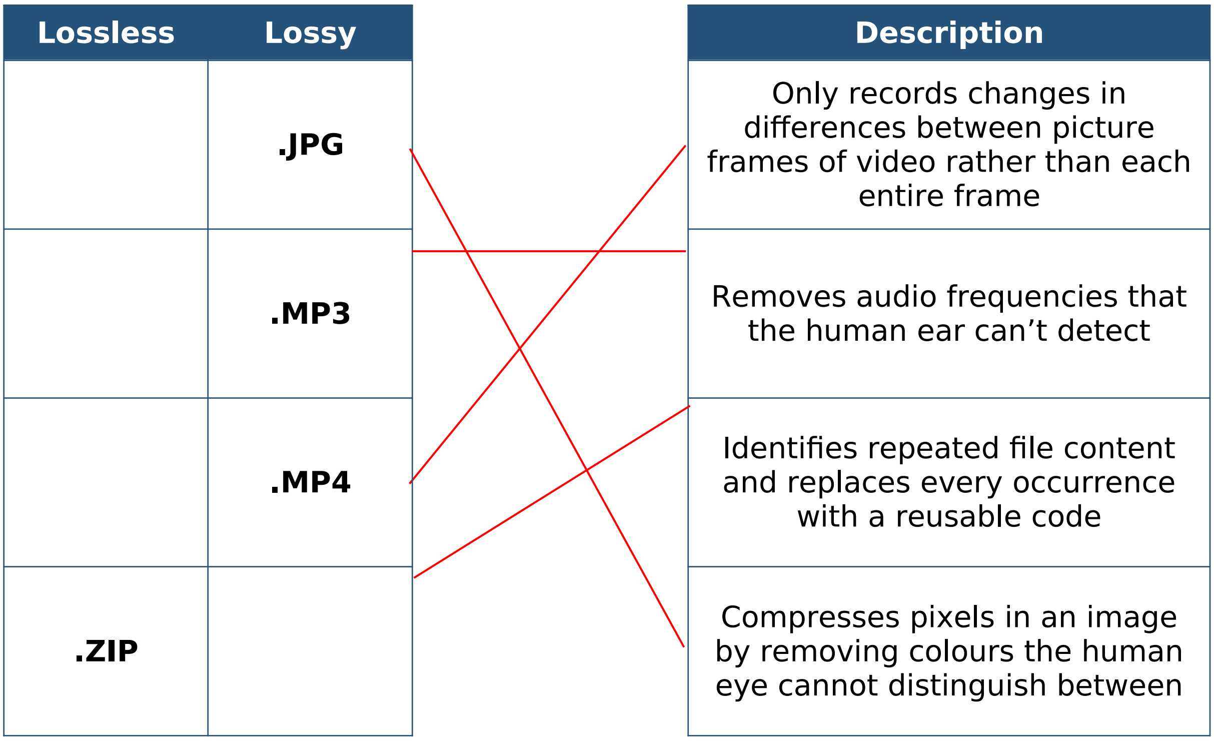
Task 1

Task 1 Compression

Files are compressed using a variety of techniques. An image file is compressed differently than a folder containing text files.

Different compression schemes produce files in a specific format. The extension of a file indicates what this is so that the Operating System opens the file with an application that can decompress it.

Complete the diagram by matching the file type with the type of compression performed, the description of what it does and the file extension.

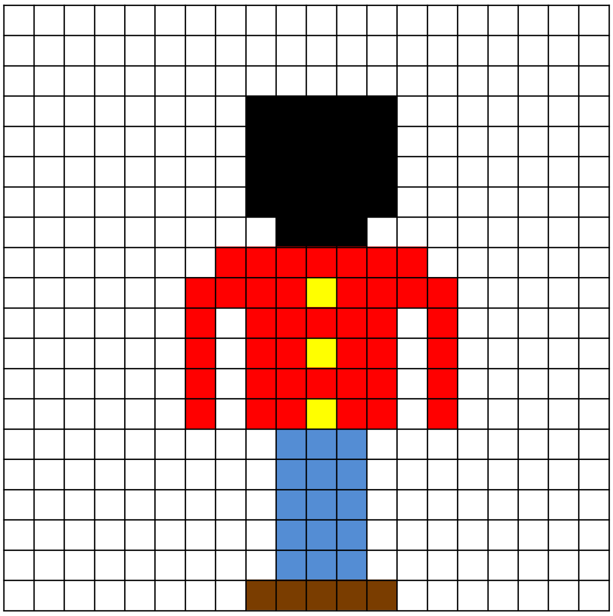


Lossless	Lossy
	.JPG
	.MP3
	.MP4
.ZIP	

Description
Only records changes in differences between picture frames of video rather than each entire frame
Removes audio frequencies that the human ear can't detect
Identifies repeated file content and replaces every occurrence with a reusable code
Compresses pixels in an image by removing colours the human eye cannot distinguish between

Task 2 Calculating compression

A 20x20 bitmap image file is described using a colour depth of 8 bits as shown below:



The file size of this image is calculated by firstly determining the resolution or how many pixels there are: 20 x 20 = 400 pixels.

Each pixel is described by 8 bits so the file size is: 400 pixels x 8 bits = 3,200 bits or 400 bytes.

a) Run length encoding is applied to the file in order to reduce its file size. An encoded sequence of the same data is stored as 2 bytes: the value of the repetition and the colour code. For example the first line would be stored as 20W (the W would be the binary code for white but is summarised here as a character).

For each line show how the data would be encoded using the format: [LengthColour]. The first and fourth lines have been done for you.

(**W**=White, **K**=Black, **R**=Red, **Y**=Yellow, **B**=Blue, **N**=Brown)

Line	Run length encoded sequence
1	20W
2	20W
3	20W
4	8W5K7W
5	8W5K7W
6	8W5K7W
7	8W5K7W
8	9W3K8W
9	7W7R6W
10	6W4R1Y4R5W
11	6W1R1W5R1W1R5W
12	6W1R1W2R1Y2R1W1R5W
13	6W1R1W5R1W1R5W
14	6W1R1W2R1Y2R1W1R5W
15	9W3B8W
16	9W3B8W
17	9W3B8W
18	9W3B8W
19	9W3B8W
20	8W5N7W

b) Determine what the RLE compressed file size is bytes:

152 bytes

c) Express, as a whole number percentage of the original file size, how much compression has been applied to the picture:

400 – 152 = 248 bytes saved
248 / 400 * 100 = 62%

d) This type of compression is categorised as lossless. Explain what is meant by this in terms of the picture file:

Lossless compression means no data is lost in the process of compressing
When uncompressed to picture will be exactly the same as the original

Task 3 Dictionary compression

A text document contains the proverb:

give a man a fish and you feed him for a day,
teach a man to fish and you feed him for a lifetime

Dictionary compression is applied to the file whereby common **single words and associated space characters** are stored in a separate dictionary component and referenced by an 8 bit binary code.

a) Each character is stored as an 8 bit ASCII code (including spaces). Calculate the storage size of the quote in **bytes**:

97 bytes

b) Complete the dictionary below listing all the repeated words in the order the proverb is presented (note that the underscore represents the space character:

Code	Repeated word
00000000	a_
00000001	man_
00000010	fish_
00000011	and_
00000100	you_
00000101	feed_
00000110	him_
00000111	for_

c) What will be the size of the dictionary in bytes if each letter of each repeated word is stored in 1 byte?

40 bytes

d) What will be the new size of the quote in bytes if codes from the dictionary are used in place of the repeated words?

44 bytes

Note that the pattern of the message would be as follows where * indicates a dictionary word has been used:

*give **********day, teach **to *******lifetime*

e) Express, as a percentage of the original quote size, how much compression has been applied to the text:

Total file size is quote + dictionary: 40 + 44 = 84 bytes
Saving from compression is 11 bytes
Percentage is 11 / 97 * 100 = 13%

f) Dictionary compression can be adjusted to be more effective by storing phrases rather than just single words.

Explain how this concept would significantly reduce the number of bytes used to represent the quote looked at in part (a).

There are two main phrases: ‘*a man* ‘, and ‘*fish and you feed him for a* ‘
This would only require three entries in the dictionary (including a separate entry for ‘*a* ‘) compared to eight previously.

Using phrases like this would result in a phrase of 33 bytes and a dictionary of 38 bytes, total 71 bytes.

Saving of 97 - 71 = 26 bytes or 27% which is more than double the previous example.

Task 4 Encryption

The Caesar cipher

Messages sent across a network are encrypted using a Caesar cipher.

a) Determine what the encrypted message of "*mary had a little lamb*" would be with a shift of minus 3:

- jxov exa x ifqqib ixjy

b) Decrypt the short message "fdwfk wkh sljhrq".

- catch the pigeon

c) The following message has been encrypted using the Caesar cipher but word length and letter cases have been obscured. Use some basic frequency analysis to decrypt the message below: (Punctuation does not shift.)

```
BRXFD QQRWO RVHDK RPLQJ SLJHR Q.LIB RXUKR PLQJS LJHRQ
GRHVQ RWFRP HEDFN ,WKHQ ZKDWB RXKDY HORVW LVDSL JHRQ.
```

The most frequently occurring of letters in the English alphabet are **E**, **T**, **A**, **O**, **I** and **N**.
The least frequent are **Z**, **Q**, **J** and **K**.

Letter	Frequency	Letter	Frequency	Letter	Frequency	Letter	Frequency
A	-	H	8	O	2	V	4
B	3	I	1	P	3	W	5
C	-	J	5	Q	9	X	3
D	6	K	5	R	14	Y	1
E	1	L	7	S	3	Z	1
F	3	M	-	T	-		
G	1	N	1	U	1		

- No As or Cs in the frequency analysis = could be X and Z
- Highest frequency is R and Q = could be N and O as they are among the top most frequent letters and also next to each other in the alphabet.
- Shift = 3
- Message = "You cannot lose a homing pigeon. If your homing pigeon does not come back, then what you have lost is a pigeon."

The Vernam cipher

The Vernam cipher has been proven to be the only cipher that is unbreakable as long as certain rules are followed. These are that the one-time pad must be truly random, used only once and must be hand delivered to the recipient.

a) Use the ASCII code sheet to encrypt the following plaintext: **Rat** with the one-time pad of: **a!H**.

b)

1	0	1	0	0	1	0	**R**
1	1	0	0	0	0	1	**a**
0	1	1	0	0	1	1	**3**

1	1	0	0	0	0	1	**a**
0	1	0	0	0	0	1	**!**
1	0	0	0	0	0	0	**@**

1	1	1	0	1	0	0	**t**
1	0	0	1	0	0	0	**H**
0	1	1	1	1	0	0	**<**

Reverse the XOR operation to decrypt the following ciphertext: **}#<** using a one-time pad of: **5L[**.

c)

1	1	1	1	1	0	1	**}**
0	1	1	0	1	0	1	**5**
1	0	0	1	0	0	0	**H**

0	1	0	0	0	1	1	**#**
1	0	0	1	1	0	0	**L**
1	1	0	1	1	1	1	**o**

0	1	1	1	1	0	0	**<**
1	0	1	1	0	1	1	**[**
1	1	0	0	1	1	1	**g**

Why must the one-time pad be generated from a truly random source rather than being computer generated?

- A computer generated random number is based on an algorithm to create it. It is therefore not actually random. If a computer created something by algorithm, another computer could replicate it using the same algorithm even if it may take decades to do it. A truly random event could never be replicated by a computer.

d) Why should a one-time pad only be used once?

- Using it a second time would create two comparative sets of cipher code that both use the same key. This would remove any randomness and having intercepted both sets would make the code not only breakable, but relatively easy to break.

e) Why must a one-time pad be hand delivered?

- Sending the key electronically would require another one-time pad to encrypt it, which would require another to encrypt that etc.
- Sending it by post or pigeon has obvious flaws.
- Hand delivering is the only certain way to ensure that the right person receives it without any risk of interception. If the key were intercepted or lost, another key would be used. It would be possible to hand deliver a set of say 100 keys at once that would be used in turn.